ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.048:378.14.015.62

АЛГОРИТМ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ, ОСНОВАННЫЙ НА МЕТОДЕ ПОЛИАДИЧЕСКИХ ЧИСЕЛ

канд. техн. наук, доц. А.Ф. ОСЬКИН, Е.В. ДАНЧЕНКО
(Полоцкий государственный университет);
Д.А. ОСЬКИН
(Белорусский государственный экономический университет, Минск)

Рассматривается полиадическая система счисления и метод полиадических чисел, использующий эту систему. Показано, что на основании метода полиадических чисел может быть построен простой, но эффективный алгоритм многомерной классификации. Приводится пример применения разработанного алгоритма для прогнозирования успешности обучения студентов первого курса с помощью многомерной классификации, в качестве исходных данных для которой используются результаты первой экзаменационной сессии.

Ключевые слова: *метод полиадических чисел, алгоритм многомерной классификации, прогноз успешности обучения.*

Введение. Методом полиадических чисел называется метод кодирования информации, использующий полиадическую систему счисления [1]. Известно, что в позиционной системе счисления с постоянным основанием всякое число $N = (x_1, x_2, \ldots, x_{m-1}, x_m)$ может быть представлено как

$$N = x_1 \cdot p^{(m-1)} + x_2 \cdot p^{(m-2)} + \cdots + x_{(m-1)} \cdot p + x_m, \quad (1)$$

где x_i – i-я цифра числа N и $x_i < p_i$, а каждый i-й разряд имеет весовой коэффициент p^{m-i}.

Если же за основание системы счисления принять не постоянное число p, а некоторый набор положительных чисел $l_1, l_2, \ldots, l_m$, на разность которых не накладывается никаких ограничений, то любое число $M = (y_1, y_2, \ldots, y_{m-1}, y_m)$ можно представить в виде

$$M = y_1 \cdot l_2 \cdot l_3 \cdot \ldots \cdot l_m + y_2 \cdot l_3 \cdot l_4 \cdot \ldots \cdot l_m + \cdots + y_{m-1} \cdot l_m + y_m, \quad (2)$$

где y_i – i-я цифра числа M и $y_i < p_i$, а каждый i-й разряд имеет весовой коэффициент $\prod_{(k=i+1)}^{m} l_k$.

Полиадическая система счисления обладает рядом интересных свойств, позволяющих в частности построить на ее основе простой, но эффективный алгоритм многомерной классификации.

Алгоритм многомерной классификации. Пусть исследуемое явление представляется в виде направленной системы данных, для которой определены свойства упорядоченности и расстояния. Будем рассматривать дискретные системы, что не ограничивает общности предлагаемых методов, но предпочтительнее по многим соображениям. Пусть в рассматриваемой системе есть единственная выходная переменная, имеющая бинарный тип и принимающая значение 1 или 0, в зависимости от того принадлежит или не принадлежит анализируемый объект данному классу объектов.

Предлагаемый алгоритм состоит из следующих шагов.

Шаг 1. Формируется исходная система данных. Выполняются процедуры извлечения, очистки и загрузки данных из различных источников. При этом в качестве источников могут выступать реляционные базы данных, файлы в формате .xml, плоские файлы, электронные таблицы и т.д.

На этапе извлечения данных выполняются их приведение к единому формату и проверка на соответствие сформулированным требованиям.

Далее данные, полученные из различных источников, объединяются, исключаются повторяющиеся записи. При необходимости данные кодируются или декодируются, удаляются явно ошибочные записи.

На основании очищенных данных формируется исходная система.

Шаг 2. Исходная система данных разбивается на две подсистемы – обучающую и проверочную.

Шаг 3. На данных обучающей последовательности проверяются гипотезы реконструкции

$$\{y_1 / y_2 / \ldots / y_m\},\ \{y_1y_2 / y_1y_3 / \ldots / y_{m-1}y_m\},\ \{y_1y_2y_3 / y_1y_2y_4 / \ldots / y_{m-2}y_{m-1}y_m\} \text{ и т.д.}, \quad (3)$$

где y_i, y_iy_j, $y_iy_jy_k$ и т.д. – полиадические числа соответственно из одной, двух, трех и т.д. цифр, кодирующие вектор характеристик объекта.

Шаг 4. Для каждой из гипотез реконструкции выполняется упорядочивание обучающей последовательности и подсчитывается суммарное расстояние между значениями выходной переменной. В качестве основной принимается гипотеза, дающая минимальное расстояние.

Шаг 5. Для этой гипотезы выделяются диапазоны изменения соответствующих полиадических чисел, на которых расстояние между значениями выходной переменной равно нулю. При этом, если выходная переменная на этом интервале принимает значение «0», то соответствующий вектор относится к классу «0», если же выходная переменная имеет значение «1», то вектор относится к классу «1».

Шаг 6. Качество принятой гипотезы контролируется на данных проверочной последовательности.

Приведем пример реализации описанного алгоритма на примере решения задачи прогнозирования успешности учебной деятельности. Аналогичная задача рассматривалась нами в работе [2]. В этой статье для построения прогноза был использован метод К–средних, реализованный в системе интеллектуального анализа данных WEKA. При этом качество прогноза проверялось путем построения ROC–кривой, выполненного с использованием надстройки AtteStat для табличного процессора MS Excel.

Применение алгоритма многомерной классификации для прогнозирования успешности учебной деятельности. Напомним, как была сформулирована решаемая задача. Имеются результаты сдачи первой экзаменационной сессии студентами факультета информационных технологий Полоцкого государственного университета. На основании этих данных требуется построить прогноз успешности окончания высшего учебного заведения в установленный учебным планом срок. Так как в качестве исходных данных были использованы сведения о студентах, принятых на факультет в 2008 году и окончивших обучение в 2013 году, имелась возможность сравнить прогноз с реальным результатом.

Исходные данные представлены в таблице 1.

В соответствии с изложенным выше алгоритмом выполним извлечение, очистку и загрузку данных. В нашем случае это достаточно простая процедура, так как источник данных всего один – электронная таблица в формате .xlsx, содержащая результаты сдачи экзаменов в первую экзаменационную сессию. Мы очистили эти данные, исключив записи с персональными кодами 200740010142 и 200740010115, т. к. студенты с этими кодами были отчислены по причинам, не связанным с их академической неуспеваемостью, и, следовательно, не должны учитываться в последующем анализе.

Теперь разобьем исходные данные на две таблицы. Первая из них будет содержать 28 записей, выбранных из исходной таблицы случайным образом. Эту выборку мы будем в дальнейшем использовать как обучающую (таблица 2).

Вторая таблица, состоящая из оставшихся 16 записей, будет использоваться нами как проверочная. Эта последовательность представлена таблицей 3.

На данных обучающей последовательности проверяем гипотезы реконструкции, используя последовательно полиадические числа, состоящие из одной, двух, трех и четырех цифр:

$$\{y_{1i} / y_{2i} / \ldots / y_{mi}\},\ \{y_{1i}y_{2i} / y_{1i}y_{3i} / \ldots / y_{(m-1)i}y_{mi}\},\ \{y_{1i}y_{2i}y_{3i} / y_{1i}y_{2i}y_{4i} / \ldots / y_{(m-2)i}y_{(m-1)i}y_{mi}\} \text{ и т.д.}, \quad (4)$$

где y_{1i} – оценка *i*-го студента по дисциплине «История Беларуси»;

y_{2i} – оценка *i*-го студента по дисциплине «Высшая математика»;

y_{3i} – оценка *i*-го студента по дисциплине «Основы алгоритмизации и программирования» (ОА и П);

y_{4i} – оценка *i*-го студента по дисциплине «Начертательная геометрия и графика» (НГ и Г).

Как показали выполненные в соответствии с шагом 4 алгоритма расчеты, наилучший результат дает полиадическое число $y_3y_2y_4y_1$.

Отсортируем теперь данные таблицы 2 в порядке убывания по новому столбцу «Полиадическое число» (ПЧ). Результат такого преобразования представлен в таблице 4.

Как видно из таблицы, уверенно успешными являются студенты, для которых полиадическое число $y_3y_2y_4y_1$, записанное в десятичной системе счисления, больше **6144**.

В нижней части таблицы общую картину «портит» студент, с персональным кодом 200740010128, который сумел закончить университет в отведенные учебным планом сроки, несмотря на серьезную неудачу в первой сессии.

Таблица 5 представляет собой отсортированные в порядке убывания по столбцу «Полиадическое число» (ПЧ) данные таблицы 3.

Как видно из таблицы, и здесь значение **6144** является пороговым, отделяющим успешных студентов от неуспешных.

При этом в верхней части таблицы 5, как и в таблице 4, имеет место стопроцентное совпадение прогноза с результатом, а в нижней части есть случаи ошибочного предсказания. Поражает пример студента с персональным кодом 200740010138, который получил в первую сессию две двойки по ключевым дисциплинам и, тем не менее, сумел завершить учебу в срок.

Таблица 1. – Итоги первой экзаменационной сессии

Персональный код студента	История Беларуси	Высшая математика	ОА и П	НГ и Г	Признак успешности
200740010100	*4*	*4*	*2*	*5*	*ОТЧ*
200740010101	*4*	*2*	*2*	*2*	*ОТЧ*
200740010102	*4*	*4*	*2*	*4*	*ОТЧ*
200740010103	*9*	*10*	*7*	*5*	*У*
200740010104	*2*	*6*	*2*	*4*	*ОТЧ*
200740010105	*7*	*6*	*7*	*6*	*У*
200740010106	*9*	*9*	*10*	*9*	*У*
200740010107	*4*	*7*	*4*	*6*	*ОТЧ*
200740010108	*6*	*4*	*5*	*4*	*ОТЧ*
200740010109	*6*	*4*	*2*	*2*	*ОТЧ*
200740010110	*4*	*6*	*5*	*4*	*ОТЧ*
200740010111	*4*	*2*	*4*	*5*	*ОТЧ*
200740010112	*9*	*7*	*7*	*4*	*У*
200740010113	*7*	*4*	*5*	*4*	*ОТЧ*
200740010114	*2*	*2*	*4*	*4*	*ОТЧ*
200740010115	*5*	*7*	*10*	*5*	*ОТЧ*
200740010116	*8*	*9*	*5*	*4*	*У*
200740010117	*8*	*6*	*5*	*6*	*У*
200740010118	*9*	*9*	*7*	*6*	*У*
200740010119	*4*	*2*	*6*	*6*	*У*
200740010120	*8*	*8*	*6*	*7*	*У*
200740010121	*6*	*7*	*4*	*4*	*ОТЧ*
200740010122	*7*	*9*	*10*	*5*	*У*
200740010123	*4*	*9*	*8*	*7*	*У*
200740010124	*8*	*9*	*9*	*7*	*У*
200740010125	*5*	*2*	*5*	*4*	*ОТЧ*
200740010126	*8*	*9*	*8*	*9*	*У*
200740010127	*6*	*6*	*6*	*5*	*У*
200740010128	*4*	*6*	*2*	*5*	*У*
200740010129	*9*	*7*	*6*	*6*	*У*
200740010130	*7*	*5*	*4*	*5*	*ОТЧ*
200740010131	*4*	*4*	*2*	*4*	*ОТЧ*
200740010132	*4*	*4*	*4*	*4*	*ОТЧ*
200740010133	*7*	*6*	*7*	*5*	*У*
200740010134	*8*	*8*	*7*	*9*	*У*
200740010135	*4*	*6*	*5*	*6*	*У*
200740010136	*5*	*2*	*4*	*4*	*У*
200740010137	*4*	*2*	*4*	*5*	*ОТЧ*
200740010138	*6*	*2*	*2*	*6*	*У*
200740010139	*6*	*4*	*2*	*6*	*У*
200740010140	*4*	*2*	*4*	*4*	*ОТЧ*
200740010141	*8*	*7*	*8*	*8*	*У*
200740010142	*6*	*6*	*8*	*8*	*ОТЧ*
200740010143	*9*	*9*	*9*	*6*	*У*
200740010144	*8*	*6*	*6*	*7*	*У*
200740010145	*9*	*6*	*4*	*7*	*У*

Таблица 2. – Обучающая последовательность

Персональный код студента	История Беларуси	Высшая математика	ОА и П	НГ и Г	Признак успешности
200740010100	4	4	2	5	ОТЧ
200740010101	4	2	2	2	ОТЧ
200740010102	4	4	2	4	ОТЧ
200740010103	9	10	7	5	У
200740010104	2	6	2	4	ОТЧ
200740010105	7	6	7	6	У
200740010107	4	7	4	6	ОТЧ
200740010108	6	4	5	4	ОТЧ
200740010109	6	4	2	2	ОТЧ
200740010110	4	6	5	4	ОТЧ
200740010111	4	2	4	5	ОТЧ
200740010113	7	4	5	4	ОТЧ
200740010116	8	9	5	4	У
200740010118	9	9	7	6	У
200740010119	4	2	6	6	У
200740010120	8	8	6	7	У
200740010122	7	9	10	5	У
200740010125	5	2	5	4	ОТЧ
200740010127	6	6	6	5	У
200740010128	4	6	2	5	У
200740010130	7	5	4	5	ОТЧ
200740010132	4	4	4	4	ОТЧ
200740010134	8	8	7	9	У
200740010135	4	6	5	6	У
200740010137	4	2	4	5	ОТЧ
200740010140	4	2	4	4	ОТЧ
200740010141	8	7	8	8	У
200740010143	9	9	9	6	У

Таблица 3. – Проверочная последовательность

Персональный код студента	История Беларуси	Высшая математика	ОА и П	НГ и Г	Признак успешности
200740010106	9	9	10	9	У
200740010112	9	7	7	4	У
200740010114	2	2	4	4	ОТЧ
200740010117	8	6	5	6	У
200740010121	6	7	4	4	ОТЧ
200740010123	4	9	8	7	У
200740010124	8	9	9	7	У
200740010126	8	9	8	9	У
200740010129	9	7	6	6	У
200740010131	4	4	2	4	ОТЧ
200740010133	7	6	7	5	У
200740010136	5	2	4	4	У
200740010138	6	2	2	6	У
200740010139	6	4	2	6	У
200740010144	8	6	6	7	У
200740010145	9	6	4	7	У

Таблица 4. – Отсортированная по столбцу «ПЧ» таблица 2

Персональный код студента	История Беларуси	Высшая математика	ОА и П	НГ и Г	Признак успешности	ПЧ
200740010122	*7*	*9*	*10*	*5*	*У*	11957
200740010143	*9*	*9*	*9*	*6*	*У*	10869
200740010141	*8*	*7*	*8*	*8*	*У*	9588
200740010103	*9*	*10*	*7*	*5*	*У*	8759
200740010118	*9*	*9*	*7*	*6*	*У*	8669
200740010134	*8*	*8*	*7*	*9*	*У*	8598
200740010105	*7*	*6*	*7*	*6*	*У*	8367
200740010120	*8*	*8*	*6*	*7*	*У*	7478
200740010127	*6*	*6*	*6*	*5*	*У*	7256
200740010119	*4*	*2*	*6*	*6*	*У*	6864
200740010116	*8*	*9*	*5*	*4*	*У*	6448
200740010135	*4*	*6*	*5*	*6*	*У*	6164
200740010110	*4*	*6*	*5*	*4*	*ОТЧ*	6144
200740010113	*7*	*4*	*5*	*4*	*ОТЧ*	5947
200740010108	*6*	*4*	*5*	*4*	*ОТЧ*	5946
200740010125	*5*	*2*	*5*	*4*	*ОТЧ*	5745
200740010107	*4*	*7*	*4*	*6*	*ОТЧ*	5164
200740010130	*7*	*5*	*4*	*5*	*ОТЧ*	4957
200740010132	*4*	*4*	*4*	*4*	*ОТЧ*	4844
200740010111	*4*	*2*	*4*	*5*	*ОТЧ*	4654
200740010137	*4*	*2*	*4*	*5*	*ОТЧ*	4654
200740010140	*4*	*2*	*4*	*4*	*ОТЧ*	4644
200740010128	*4*	*6*	*2*	*5*	*У*	2854
200740010104	*2*	*6*	*2*	*4*	*ОТЧ*	2842
200740010100	*4*	*4*	*2*	*5*	*ОТЧ*	2654
200740010102	*4*	*4*	*2*	*4*	*ОТЧ*	2644
200740010109	*6*	*4*	*2*	*2*	*ОТЧ*	2626
200740010101	*4*	*2*	*2*	*2*	*ОТЧ*	2424

Таблица 5. – Отсортированная по столбцу «ПЧ» таблица 3

Персональный код студента	История Беларуси	Высшая математика	ОА и П	НГ и Г	Признак успешности	ПЧ
200740010106	*9*	*9*	*10*	*9*	*У*	11999
200740010124	*8*	*9*	*9*	*7*	*У*	10878
200740010126	*8*	*9*	*8*	*9*	*У*	9798
200740010123	*4*	*9*	*8*	*7*	*У*	9774
200740010112	*9*	*7*	*7*	*4*	*У*	8449
200740010133	*7*	*6*	*7*	*5*	*У*	8357
200740010129	*9*	*7*	*6*	*6*	*У*	7369
200740010144	*8*	*6*	*6*	*7*	*У*	7278
200740010117	*8*	*6*	*5*	*6*	*У*	6168
200740010121	*6*	*7*	*4*	*4*	*ОТЧ*	5146
200740010145	*9*	*6*	*4*	*7*	*У*	5079
200740010136	*5*	*2*	*4*	*4*	*У*	4645
200740010114	*2*	*2*	*4*	*4*	*ОТЧ*	4642
200740010139	*6*	*4*	*2*	*6*	*У*	2666
200740010131	*4*	*4*	*2*	*4*	*ОТЧ*	2644
200740010138	*6*	*2*	*2*	*6*	*У*	2466

Заключение. Проведенные исследования позволяют сделать следующие выводы:

1. Метод полиадических чисел, использующий полиадическую систему счисления позволяет построить простую, но эффективную систему многомерной классификации.

2. Использование предложенного метода для решения задачи прогнозирования позволило построить достаточно точный прогноз успешности обучения для группы студентов, получивших в первую экзаменационную сессию по ключевым предметам («ОА и П» и «Высшая математика») оценки не ниже шести баллов. Это хорошо согласуется с практикой оценивания компетенций, распространенной в странах, подписавших Болонское соглашение. В ряде стран-участниц Болонского процесса кредиты за пройденную дисциплину, входящую в список ключевых для данной специальности, засчитываются только в том случае, если оценка по ней в пересчете на белорусскую систему оценивания знаний не меньше шести баллов.

3. Прогноз оказался менее точным в части касающейся неуспешных студентов. Как уже указывалось выше, в группу успешных попал студент с персональным кодом 200740010138, чего не должно было быть, так как этот студент получил две двойки по ключевым дисциплинам – «ОА и П» и «Высшая математика».

4. Тем не менее, предложенный метод прогнозирования успешности академической деятельности может быть рекомендован к внедрению в деканатах высших учебных заведений, так как позволяет деканам, заведующим кафедрами, ведущим преподавателям сфокусировать свое внимание на неблагополучных студентах и помочь им справится с возникшими проблемами.

ЛИТЕРАТУРА

1. Амелькин, В.А. Методы нумерационного кодирования / В.А. Амелькин. – Новосибирск : Наука : Сиб. отд-е, 1986. – 155 с.
2. Оськин, А.Ф. Применение интеллектуального анализа образовательных данных для прогнозирования успешности учебной деятельности / А.Ф. Оськин, Д.А. Оськин // Вестн. Полоц. гос. ун-та. Сер. С, Фундам. науки. – 2016. – № 4. – С. 8–12.

Поступила 03.09.2019

THE ALGORITHM FOR MULTIDIMENSIONAL CLASSIFICATION BASED ON THE POLYADIC NUMBERS

A. OSKIN, E. DANCHENKO, D. OSKIN

The polyadic number system and the method of polyadic numbers using this system are considered. It is shown that on the basis of the method of polyadic numbers a simple but effective algorithm of multidimensional classification can be constructed. An example of application of algorithm for prediction of academic success of first-year students using multi-dimensional classification, as the initial data for which the results of the first examination session.

Keywords: *method of polyadic numbers, algorithm of multidimensional classification, prediction of learning success.*